数理解析研究所講究録
第790巻 1992年 22-28

決定木構成問題と学習可能性

On the Decision Trees: Construction Problem and Learnability

小柴健史 (Takeshi Koshiba)
東京工業大学 理工学研究科 情報工学専攻

1 はじめに

決定木とは，与えられた対象の集まりを各対象が持つ性質を利用して分類する1つの方法である．その簡便性も手伝って決定木は計算機科学の分野はいうに及ばず，生物学などの自然科学分野においても広く利用されている．計算機科学分野においては，データベース，パターン認識，機械による自動判定・認証，論理回路設計，アルゴリズムの解析等に利用されている [Mor82]．こうした応用では多くの場合，“小さな” 決定木が望まれるが，一般に最適化問題は計算論的に難しく，最小決定木構成問題も NP 困難であることが示されている [KW91, Han89]．しかし，最小の決定木でなくても十分小さな決定木が構成できればかなり有用なことが多く，近似的に小さな決定木を求めるヒューリスティックな手法が色々と研究されている [Qui86, Utg89, Mor82]．本稿では，そうしたヒューリスティックな手法を理論的に解析し，その可能性や限界を示す．また，決定木の学習という立場から「小さな決定木を構成すること」に対して理論的な意味を与える．

2 決定木について

決定木 (decision tree) とは，与えられた事例 (example) の集合を，各事例に付随する属性 (attribute) の値によりいくつかの集合に分類する木のことである．

与えられる事例に対して，各々 m 個の属性値 (0 または 1) と，その分類値 (Y または N) が定義されている．このような事例の集合をサンプルと呼ぶ．たとえば表 1 に示すような事例 $e_1, \ldots, e_5$ の集合 S がサンプルである．サンプルのサイズとはサンプル内の事例の数である．表 1 のサンプルのサイズは 5 である．

	属性値							分類値
事例	x_1	x_2	x_3	x_4	x_5	x_6	x_7	f
e_1	1	1	0	0	1	0	1	Y
e_2	1	1	1	1	0	0	0	N
e_3	0	1	0	0	1	1	0	N
e_4	1	0	0	0	0	1	1	Y
e_5	0	0	1	0	1	1	1	N

$S = \{e_1, e_2, e_3, e_4, e_5\}$

表 1: サンプルの一例

これに対し，各事例の属性値をもとに事例を分類値が Y のものと，N のものへ分類する一

手法が決定木である．たとえば図 1 の木が決定木の一例だが，一般に**決定木**とは，内部頂点には属性名が，葉には Y または N が，それぞれラベル付けされた二分木のことである．

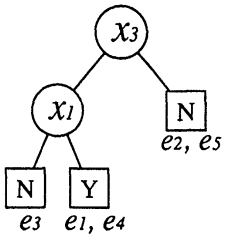

図 1: 決定木の一例

各事例には，二分木の根から出発し葉へ到達する道（path）がユニークに対応する．つまり，内部頂点にラベル付けされた属性に対し，その事例の属性値が 0 ならば左へ，1 ならば右へ，と進む道である．たとえば先の事例 $e_1, \ldots, e_5$ は，それぞれ図 1 の決定木をその対応する道に沿って進むと，図中に示す葉に到達する．サンプル中のすべての事例に対し，その到達先の葉のラベルが事例の分類値と一致する場合に，その二分木はサンプルに対して**無矛盾**であるという．たとえば図 1 の木は先のサンプル S に無矛盾な決定木になっている．

ところで，決定木の大きさの尺度はいろいろあり，その尺度により “小さな決定木” の意味するところも変わってくる．本稿では木 T の大きさの尺度として木の葉数を採用し，$\sigma(T)$ で表す．

3 コンパクトな決定木と学習可能性

学習可能性とデータ圧縮との関連の理論的な裏付けとしてオッカムアルゴリズムとの関係が知られている [BEHW87]．この結果を利用して本節では小さな決定木を得ることの意味を計算論的学習理論の立場から与える．(オッカムアルゴリズムの定義は [BEHW87] を参照のこと)．

一般の学習可能性についての議論を決定木の学習について適用するための準備をする．

補題 3.1. a を属性数とする．高々葉数が ℓ の決定木の数を $r(\ell)$ で表す．このとき，$r(\ell) < (8a)^\ell$ である．

証明. 葉数が i のとき木の総数は $\frac{1}{2i-1}\begin{pmatrix} 2i-1 \\ i \end{pmatrix}$ なので，

$$\sum_{i=1}^{\ell} \frac{2^i a^{i-1}}{2i-1}\begin{pmatrix} 2i-1 \\ i \end{pmatrix} \;<\; (2a)^\ell \sum_{i=1}^{\ell}\begin{pmatrix} 2\ell-1 \\ i \end{pmatrix} \;<\; (2a)^\ell 2^{2\ell-1} \;<\; (8a)^\ell$$

□

上の補題を [BEHW87] の結果に適用し決定木の大きさの観点で換言すると以下のようなことがいえる．

定理 3.2. m をサンプルサイズ，h_0 を出力仮説の表現の大きさの上限とする．

$$\left(\frac{2\ell_0^c \ln 8a}{-\ln(1-\varepsilon)}\right)^{\frac{1}{1-\alpha}} > \frac{2\ln(1/\delta)}{-\ln(1-\varepsilon)}$$

のとき,

$$h_0 \leq \frac{m\varepsilon + \ln \delta}{\ln(8a)}$$

なる関係があれば,

$$\Pr[\text{ オッカムアルゴリズムの出力 } h \text{ に対して } \Pr[\ f(e) \neq h(e) \mid e \in_P X\ ] < \varepsilon\ ] \\ \mid \text{サイズ } m \text{ のサンプル } S \in_{Pm} X^m] \geq 1 - \delta$$

を満たす.

このことより, 十分小さな決定木を求めるアルゴリズムが存在すれば, 得られた決定木は未知の事例に対して誤って分類してしまう確率が低い値で抑えられることが保証される.

4 ローカル戦略クラス

Quinlan [Qui86] の決定木学習システム ID3 は, あるアルゴリズム — 以下ではこれを Quinlan のアルゴリズムと呼ぶ — に従ってサンプルに無矛盾な決定木を構成している. ここでは, その構成方法で果たして最小決定木が作れるか? あるいはどの程度小さな決定木が作れるか? について調べる. ただし, 我々は Quinlan のアルゴリズムだけを対象に解析を行なうのではない. 実は, Quinlan のアルゴリズムは, かなり一般的な方針 — ローカル戦略 — に従ったアルゴリズムの一つとみなせる. ここでの解析結果は, こうしたローカル戦略に従うアルゴリズムすべてに成り立つ結果である.

4.1 ローカル戦略とその例

定義 4.1. いま考えている属性を $x_1, x_2, \ldots, x_n$ とし, これらの集合を V とする. 事例の全体集合を X とし, サンプルを $S \subseteq X$ とする. **評価関数** g を V と 2^X の直積空間からある順序集合への写像とする. まず, サンプル S を 属性 x_i により 4 つに分割をする.

$$S_N^0(x_i) = \{e \in S : x_i(e) = 0, f(e) = \mathrm{N}\}, \quad S_N^1(x_i) = \{e \in S : x_i(e) = 1, f(e) = \mathrm{N}\}$$
$$S_Y^0(x_i) = \{e \in S : x_i(e) = 0, f(e) = \mathrm{Y}\}, \quad S_Y^1(x_i) = \{e \in S : x_i(e) = 1, f(e) = \mathrm{Y}\}$$

これらの分割を用いて,

$$g(x_i, S) \equiv \varphi(S_N^0(x_i), S_N^1(x_i), S_Y^0(x_i), S_Y^1(x_i))$$

と定義する. 評価関数 g の値を**評価値**と呼ぶ. 評価関数 g を用いて, 各属性の評価値から最適な評価値を持つ属性 x_j を選択し, ノードのラベルとする. サンプル S を 属性 x_j の属性値によって二分し, $x_j(e) = 0$ となる事例の集合を S_0, $x_j(e) = 1$ となる事例の集合を S_1 とする. これらの部分サンプル S_0 と S_1 にそれぞれ関数 g を適用して — というようにルートから順に, 木を再構成することなく決定木を構成していく. トップダウンで決定木を構成していく. このような戦略のクラスを**ローカル戦略クラス**という.

補注. S の 4 つの分割のそれぞれの要素数が等しくなるサンプル S' を考えてみる．評価関数 g を計算するのに x_i の属性値と分類値のみしか利用できないので，サンプル S と S' の違いは認識できないことになる．つまり，属性 x_i は

$$g(x_i, S) \equiv \psi(|S_N^0(x_i)|, |S_N^1(x_i)|, |S_Y^0(x_i)|, |S_Y^1(x_i)|)$$

を満たす関数 ψ で評価されることになる．ここで，一般性を失うことなく g のレンジを $\mathbf{R}$ とすることができる．以降，ψ と φ を同一視することにする．

ローカル戦略クラスのメンバーの例として，前述した Quinlan のアルゴリズムを示す．

例 4.1. いま属性 x_i がルートとしてラベル付けされる属性として選択され，属性 x_i が S を 2 つの集合 $S^0(x_i)$ と $S^1(x_i)$ に分割したとする．評価関数は

$$qnln(x_i) = -\sum_{a\in\{0,1\}} \frac{1}{y+n}\left(y_a \log \frac{y_a}{y_a+n_a} + n_a \log \frac{n_a}{y_a+n_a}\right)$$

で与えられ，この値が最小である属性を選ぶという手法が Quinlan の方法である．

ローカル戦略クラスに属するのもう一つのメンバーの例を示す．

例 4.2. 与えられたサンプル S を x_i の属性値で 2 分割する．各部分サンプル $S^a(x_i)$ において，y_a と n_a のうち少ない方に対応する事例を**誤分類**された事例という．誤分類をなるべく少なくするように選択する．つまり，属性を選択する基準として

$$wmis(x_i) = \sum_{a\in\{0,1\}} \frac{y_a+n_a}{y+n} \min\{y_a, n_a\}$$

この値を最小にする属性を選択する手法を Weighted Error 法という．

4.2 ローカル戦略の性質

我々はローカル戦略クラスに属する戦略でどの程度小さな決定木を作れるかに関心がある．次の定理はその限界を示している．

定理 4.2. ある種のサンプルに対しては，ローカル戦略クラスに属するどんな戦略を用いても最小の決定木を構成できない．

証明. [KW91] を参照． □

4.3 各ローカル戦略の評価

ローカル戦略クラスに属する任意の戦略について，全般的な性質を評価したが，ローカル戦略クラスの具体例 — Quinlan 法，Weighted Error 法 — について評価する．一般的な評価は難しいのでサンプルを限定して評価する．本稿では，次の二通りのサンプルについて評価している：(1) 属性と分類値が独立なサンプル，(2) 属性間の相関が強いサンプル．

まず，サンプル内の属性が分類値に独立であることを仮定した場合について評価する．任意の属性 x_i において分割 $S^0(x_i)$ と $S^1(x_i)$ で両者とも分類値が Y の事例数と N の事例数との比が一定であるとする．

- Weighted Error 法の場合．y, n をそれぞれ Y の事例数，N の事例数とし，y_0, y_1, n_0, n_1 をそれぞれ $S^0_Y(x_i)$，$S^1_Y(x_i)$，$S^0_N(x_i)$，$S^1_N(x_i)$ の要素数とする．また，$k = (y+n)/\min\{y, n\}$ とする．このとき，属性 x_i の評価値は

$$\begin{aligned} wmis(x_i) &= \min\{n_0, y_0\} \cdot \frac{y_0 + n_0}{y+n} + \min\{n_1, y_1\} \cdot \frac{y_1 + n_1}{y+n} \\ &= \frac{2k}{y+n}\left(\left(\min\{y_0, n_0\} - \frac{y+n}{2k}\right)^2 + \frac{(y+n)^2}{4k^2}\right) \end{aligned}$$

  のようになる．このことよりサンプルをなるべく二等分する属性が選択される．

- Quinlan 法の場合．属性 x_i の評価値は

$$\begin{aligned} qnln(x_i) &= -\frac{1}{y+n}\left(n_0 \log \frac{n_0}{y_0+n_0} + y_0 \log \frac{y_0}{y_0+n_0} + n_1 \log \frac{n_1}{y_1+n_1} + y_1 \log \frac{y_1}{y_1+n_1}\right) \\ &= -\frac{n}{y+n} \log \frac{n}{y+n} - \frac{y}{y+n} \log \frac{y}{y+n} \end{aligned}$$

  のようになる．このことよりすべての属性に対する評価値が等しいのでどの属性を選択してもよいことになる．

属性と分類値が独立なサンプルの場合はサンプルを二等分していく属性を選択するときに最小決定木を構成する．つまり，Weighted Error 法がこの場合には最適な戦略である．

次に，属性間の相関が強いサンプルの例として部分サンプル $S^0(x_i)$ に全順序がついているサンプルを考える．とくに，$i < j \leftrightarrow S^0(x_i) \subset S^0(x_j)$ を仮定しても一般性を失わない．区間とは，$S^0(x_i) \backslash S^0(x_j)$，$i > j$ で表現される部分サンプルのうちそれに属するすべての事例の分類値が同一であるような極大な事例集合のことをいう．また，区間の幅とは区間に属する事例数のことをいう．さらに，順番に並んだ事例の分類値は Y，Y，N，N，Y，Y，N，N，... のように 2 つずつ Y と N を繰り返すような幅 2 の区間サンプルを対象に評価する．

まず，このサンプルの性質について考える．部分サンプルに順序が付いているので定義から属性にも順序がついているといえる．その順序にしたがって属性の列 $\{x_i\}$ を考えることができ，属性を横軸にとり評価関数のグラフを表現できる．以下，簡単のため $M(x) = x \log x$ とする．

- 一般的な区間サンプルについて Quinlan 法を評価してみる．いま，仮に x_t で分割される点の付近で分類値が Y が続いているとする．部分サンプル $S^0(x_t)$ の Y，N の要素数を y_0, n_0 とし，部分サンプル $S^1(x_t)$ の Y，N の要素数を y_1, n_1 とする．また，s をサンプルの要素数とする．x_t 近辺で Y が続いているとすると

$$qnln(x_{t+i}) = -\frac{t+i}{s}\left(M(\frac{y_0+i}{t+i}) + M(\frac{n_0}{t+i})\right) - \frac{s-t-i}{s}\left(M(\frac{y_1-i}{s-t-i}) + M(\frac{n_1}{s-t-i})\right)$$

これを i について微分すると

$$\frac{\partial qnln(x_{t+i})}{\partial i} = \frac{1}{s}\log\frac{(y_0+n_0+i)(y_1-i)}{(y_0+i)(y_1+n_1-i)}$$

よって，この評価関数は上に凸な関数である．つまり，極小値は Y と N の境界に限定されることになり，区間の境界に相当する属性が必ず選択される．

一般の区間サンプルで区間境界に相当する属性が選択されることが分かったので幅 2 の区間サンプルでもこのことは成立する．つまり，選択される属性の候補としては $\{x_{2i}\}$ に限定できる．このときの各属性の評価値を計算する．(ただし区間数を n とする)．例として n が奇数，かつ $\{x_{4i}\}$ についての評価関数を示す．

$$\begin{aligned} qnln(x_{4i}) &= -\frac{2i}{2n}\left(M(\frac{i}{2i}) + M(\frac{i}{2i})\right) - \frac{2n-2i}{2n}\left(M(\frac{n-i+1}{2n-2i}) + M(\frac{n-i-1}{2n-2i})\right) \\ &= \frac{i}{n} - \frac{n-i}{n}\left(M(\frac{n-i+1}{2n-2i}) + M(\frac{n-i-1}{2n-2i})\right) \end{aligned}$$

属性の部分列ごとに評価関数は変化するが，いずれの場合も上に凸で山状な関数である．つまり，いずれの場合も最初のインターバルの境界に当たる属性が選択されることになる．

- 次に Weighted Error 法を考える．Weighted Error 法の場合 Quinlan 法のような区間を細分しないという（小さな木のために）よい性質はない．Weighted Error 法の評価関数を区間数 $n = 4k, k \in \mathbf{N}$，かつ $\{x_{4i}\}$ の場合について示す．

$$\begin{aligned} wmis(x_{4i}) &= \frac{1}{n}((4i)(2i) + (8k-4i)(4k-2i)) \\ &= \frac{16}{n}\left((i-k)^2 + k^2\right) \end{aligned}$$

属性の部分列ごとに評価関数は変化するが，いずれの場合も下に凸で谷状な関数である．しかも，評価値の軸と平行な直線に関して線対称な関数である．つまり，いずれの場合も極小値は $\{x_i\}$ の中央付近の属性であり，この属性が選択されることになる．

幅 2 の区間サンプルにおいて Quinlan 法と Weighted Error 法で決定木の大きさを評価する．n を区間数とする．

- Quinlan 法では常に区間境界に相当する属性を選択するので $\sigma = n$ である．
- Weighted Error 法では区間を細分してしまう属性を選択することがある．詳細は紙面の都合上省略するが，$n \leq \sigma < (3n+1)/2$ となる．

5 おわりに

本稿では木の大きさの尺度として葉数を採用して評価したが，木の大きさの尺度として総パス長（ルートから各葉までのパス長の総和）で評価すると幅 2 の区間サンプルでは Weighted Error 法の方が Quinlan 法より良い結果が得られる．

いずれの尺度からの評価において小さな決定木を得るための評価関数として，Quinlan 法のように分類値が同一であるような部分サンプルは細分せず，しかも，下に凸で谷状の関数が存在すればそれは理想的な評価関数であると推測できる．

謝辞

日頃から熱心な御指導を賜わっている東京工業大学工学部情報工学科・渡辺治助教授に深く感謝致します．

参考文献

[BEHW87] Anselm Blumer, Andrzej Ehrenfeucht, David Haussler, and Manfred K. Warmuth. Occam's razor. *Information Processing Letters* **24**(6), pp. 377–380, 1987.

[EH89] Andrzej Ehrenfeucht and David Haussler. Learning decision trees from random examples. *Information and Computation* **82**(3), pp. 231–246, 1989.

[Han89] T. Hancock. Finding the smallest consistent decision tree is NP-hard. Harvard University, 1989. Unpublished Manuscript.

[Kos92] 小柴健史. 決定木構成問題における局所的手法の性質とその可能性. 東京工業大学修士論文, 1992.

[KW91] 小柴健史, 渡辺治. 決定木の構成について. 信学技報 **91**(268), pp. 1–8, 1991.

[Mor82] Bernard M. E. Moret. Decision trees and diagrams. *Computing Surveys* **14**(4), pp. 593–623, 1982.

[Qui86] J. Ross Quinlan. Induction of decision trees. *Machine Learning* **1**(1), pp. 81–106, 1986.

[Sak91] 榊原康文. 決定木による分類規則の学習について – 理論的側面から –. 情報学基礎研究会資料, 1991.

[Utg89] Paul E. Utgoff. Incremental induction of decision trees. *Machine Learning* **4**(2), pp. 161–186, 1989.